關於統一碼 (Unicode) 漢字編碼框架的理論問題

朱　一星

〈要約〉

近年，筆者は漢字を記号論的に捉え直す試みのなかで，漢字単位説を打ち立てたが，その考えの根幹に対して有効な反論が現れない限り，デジタル情報交換用漢字の現行符号化モデルに大きな疑念を生じてもおかしくなかろう。なぜなら，「漢字単位」と称する抽象概念が，文字の形状に左右される根拠はどこにもないからである。

現行国際漢字符号化モデルは，控えめに言っても「符号であり，字形ではない（Characters，not glyphs）」という基本原則に則っているとは言いがたい。離散型の符号位置に，非離散型の異形漢字を割り当てる場合の，情報交換の伝達精度を危うくする危険性は誰もが理解できるにもかかわらず，この基本中の基本は，長いあいだ，議論の枠外に置き去りにされた模様である。

ユニコードは漢字統合をしたために不具合を生じたと俗に認識されがちだが，とんでもない誤解である。いわゆる「統合」は実質的には「同期」であり，そもそも正しい意味の「同期」を実現したことは一度もないのが事実である。それに先立つ漢字単位の同期に関する議論は，漢字研究分野で未だ注目されていない。

前世紀と違って，漢字異体字セレクタ（IVS）技術が実装できる２１世紀では，信頼に値する漢字符号化モデルを構築するためのインフラ整備について議論を深化させるべきだと筆者は主張したい。そのためには，漢字の形状に執着するあまり異体・異形字をプレーンテキストにまで押し込む愚を指摘できるユニコーダーが，今後増えていくよう切に望む。

〈キーワード〉

汉字单位　汉字编码　中日韩越统一汉字　源码分离

当我们立足于现代符号理论的原理，共时地、系统地考察汉字，便能推导出“汉字单位”概念（参见朱一星2013）。同样，我们从汉字单位的观点出发审视国际汉字编码字符集，就会看到两个突出问题。第一，目前的汉字编码常常将简化字和繁体字，以及其他各种异体字甚至讹误汉字，几乎都放在同一个层面上来对待。第二，是汉字在不同地区之间呈现出严重的理论值非同步现象。

也就是说，号称“统一”的中日韩越统一汉字编码，在框架结构上并非符合汉字符号理论意义上的同一性要求。这是因为，汉字单位概念的基本属性之一是其抽象性，体现为汉字单位（或理解为汉字的理论值）不受任何图形性要素的左右。中日韩越统一汉字虽然在一定范围内承认各国各地区汉字的笔形差异，却仍然未能摆脱汉字图形性要素的影响，决定汉字码位的基本前提，始终受到

字体或字形的摆布。

这样的编码模式如果只是在某个封闭单一的网络空间使用，也许问题还不大。事实上，汉字进入电子计算机的初期，人们也往往将其作为打字机似的工具用来打印文件。然而数码化文字的革命性意义却不仅仅是为了把文章印刷得精美绝伦，而是让文字（数码信号）往来于不同规范标准、不同语言文化背景的用户平台。一方面用于不同的显示要求，同时还需要承受数据检索、且需要经得起多次利用、更需要用于从前不可想象的巨量高速信息分析取舍。换言之，对于数码信息来说，文字的图形性特征已降为次要问题，而文字作为电磁信号的唯一性才是首要的问题。比如，对于任何一位中文阅读者来说，“经济”就等于“經濟”；“图书”无非是“圖書”；“广州”理应是“廣州”；“横滨”与“橫濱”之间也必须是等值关系。

这，就是汉字符号的一个理所当然的总原则，更是在数码信息时代必须得以体现的基本规格。

一. 焦点问题

说实话，指出上述的问题现象其实并不困难，事实上这也是笔者第一次看到的统一码标准文件时便油然而生的疑问。在统一码诞生前后，这也是信息领域一些具有深谋远虑的信息科学家及一些工程技术人员努力的方向。可是遗憾的是，实现上述原则的方法和途径就不太好办了。笔者多年来追踪考察这一问题，始终感觉如若步入五里雾中，找不到解决问题的突破口。

如今，笔者认为最为关键的焦点问题，是因为汉字从有形的传统文字时代，正在步入无形的数码文字时代。东亚地区汉字系统需要整合的要求，在上世纪70年代以前，原本并不是那么迫切的，可是到了电脑时代，尤其是网络时代，这一问题才突然变得十分突出，且颇为复杂。

有关这方面的详细思索和讨论，笔者已经另外撰文深入考究，拟刊载于香港中国语文学会的《语文建设通讯》近期号上。本文针对一些关键性概念，措辞，以及现象作一些相关的分析讨论。

对于常人来说，或许还很难想象在信息网络时代，汉字符号已经成为一个广域性的，国际性的大问题。一方面，传统的汉字理据观、正字规范、或文化审美意识，极其容易让人们产生对汉字字体字形的执着和依恋。

这样，在汉字编码问题上就会表现出多重误区：首先是各个汉字使用地区分别提出各自的编码主张，然后国际编码组织经过困难的协调达到一个看似满足各自要求的“统一的”规格。这一具有约束性的编码规格又构成对各地区汉字体系使用上的限制，也造成许多不便，甚至错误。

处在这样一个怪圈之内，人们看不见问题的根源所在，造成所谓“牵一发而动全身”的僵持局面。加上编码技术人员的刻苦努力，会在某种程度上缓解一些系统性的、危机性的漏洞，这就更使得并不十分专业的汉字使用者无法获得判断问题的依据，找不到解决问题的出发点了。

笔者将不厌其烦地表明：不管人们的主观意愿如何，网络时代的汉字系统已经不同于纸张上的汉字。标示语义的汉字符号一旦进入全球网络，汉字符号就自然而然地成为一个全球现象。汉字作为工具性的文字符号，它的本体规划和系统完善就不可能由各个地区分别承担。如此看来，如果汉字研究理论界找不到解决问题的突破口，肩负起对世界“汉字系统”提出整体技术要求的责任的话，在现时的汉字编码体制内，汉字编码问题就会始终陷在一个没有终解的泥潭之中。

二. 相关概念

为了尽量能够把问题的所在表明清晰，为了理解数码汉字符号系统性。我们有必要首先关注一些相关概念的含义以及表述概念的用词。名词术语的问题事关讨论的精确度，这一领域的许多词语概念检验工作甚至需要从零起点做起。

［**字位**］自从1990年周胜鸿提出“书同文”，呼吁解决“部分简化字和繁体字非对应”，香港中国语文学会主办的《语文建设通讯》也在1999年前后开始由姚德怀，胡百华等不断提出必须消除两岸汉字不对应，提出“**和谐体**”的用字方案。并开始在许多文章中使用了“**字位**”的概念。

“字位”一词据认为是心下（Xieyan HINCHA）首先在1985年的论文里提出的，对此无人提出异议。多位学者在说明“字位”时，包括心下本人也往往使用语音学的“音位”概念来进行类比，让读者通过对音位一词的理解去把握文字符号领域的同类性质观念。虽然这一术语尚未见到在文字学理论的语境中被理论性地加以描述，但考虑到音位概念在音位学领域具有深厚的理论基础并已经得到学界广泛一致的认可，笔者由此对“字位”做合理推测，认为字位的概念在理论上应该具有共时性和离散性。由此，笔者基本上把“字位”视为和“汉字单位”完全同等的概念。

然而，我们目前暂时还无法使用同样判断，将上述思路适用到文字理论上无法成立的“字素”一词上。如果人们用“音位”和“音素”这两个相对的概念来揣测“字位”和“字素”的定义，可能就要吃大苦头了。况且“字素”一词已出现有过其他定义，所以笔者竭力避开“字素”这一极易引起混乱的说法。

深入研究汉字和字母文字的本质性不同，笔者进一步认为：主要功能取向对应语言第一层切分的汉字既不能等同于语言系统，也不是“符号的符号”，而极有可能是并列于语言系统的另一个符号系统。笔者试将其称作“**类语言**”符号系统。而所有以标音（表音）为主要功能的音素文字和音节文字，都属于对应语言第二层切分的、细小到“不足以”标示语意的文字符号，这些文字系统应该统称为“**语音文字**”。如果这些观点符合客观事实，那么我们就必须谨慎地在分清语言和文字，汉字和语音文字本质不同的前提下讨论“字位”一词的原始定义了。但尽管如此，我们还是必须肯定，字位一词在将汉字作为符号系统考察的研究过程中做出了突破性贡献。

［**grapheme**］上世纪中叶还出现了一个文字学术语 grapheme，词典上[1)] 的一个解释是指某一音位对应字母文字时不一定唯一的书面表达，比如 /k/ 这一英语语音的文字表达可能有三种： c；k；q。grapheme 的另一个意思是指同一组文字字形的抽象概念，比如拉丁文字的大小写、印刷体和手写体等。《语言与语言学词典》（上海辞书出版社）将这一术语译作“字位”[2)]。由于这个术语的原始定义在中国文字理论领域没有系统的阐释，故将这一术语套用在汉字系统上就仍有不确定因素。今后是否能成为现代汉字符号理论研究有价值的术语还有待学界的专家进一步讨论。

［**字种**］许多学者还习惯使用“字种”的说法。据称，这一说法源于一份日本文部省1968年刊行的字表文件名称《各种汉字表・字种一览》[3)]。一般认为“字种”包括正字以外的异体字、俗字、古字，不论繁简。但是对于是否还包括各种书体（风格体），学界似乎尚未形成共识。至于中文汉字和日本汉字的同一个汉字单位是否能用“字种”来描述，目前更是无人加以明确化，这就恐怕足以令年轻学者们无所适从。比如有的学者认为“娘”虽然在中文和日文里字体是相同的，但因

为语义所指有歧义，在某种意义上不能说是同一个字种[4]。以笔者力所能及的查阅范围来看，尚还未见有关“字种”一词在符号学共时性方面的理论阐述，反而常常让人觉得只是个统计字频字数时表示“不同”汉字的量词[5]。但是问题是，汉字符号理论更加关注的是“如何不同”。恰恰在这一关键性问题上，“字种”一词似乎还没有准备好做出回答。故其做为现代汉字理论的术语性，还有待进一步的探讨和表述上的检验。

［**系统**］常常有文章对汉字的字形系列使用“系统”的说法，比如在言及简化字或繁体字的不同时，把这两个概念称为“简化字系统”和“繁体字系统”。此类“系统”，潜伏着混淆“汉字符号系统”概念的危险性。关于简化字和繁体字含义的界定，许多学者已经做过详细的说明[6]，非繁非简汉字（又称传承字）事实上占了《通用规范汉字表》的多数，这一大批汉字不归属繁体或简体，却分明属于规范所及的范围之内。另外，大家都知道大陆的规范标准还包括新字形，许多新字形的运笔设计甚至影响到笔画数，比如象“差”“骨”。新字形的覆盖面不仅和简化字不重合，也超越了非繁非简汉字，影响到印刷古籍时使用的繁体选用字，例如 體（注意有别于 體）。在大陆内地，以上所有这些无法称作简化字的字群，都可以认为是“规范”汉字。

［**正字**或**正体字**］在日本的官方文件中，习惯称经过规范的汉字为“日本正字”。如果现代汉字符号理论界认可经官方规范认可的正统汉字就是正字，那么目前无疑在除了台湾正字（正体字）以外，同时还存在大陆正字和日本正字。如果说“规范汉字”的范围有可能超过正字范围（比如大陆的繁体选用字；台湾的标准行书[7]；或日本的人名用汉字表中的异体字）。那么正字就可能和“规范汉字”不完全吻合了。

目前的编码框架所容纳的汉字表较杂乱，欠缺符号系统意义的理论性。本文尝试粗略地描述目前的汉字编码表的标量值和理想的汉字单位如何错位。为行文方便，本文称大陆的编码汉字表为GB汉字。台湾的编码汉字为Big5汉字[8]，日本汉字为JIS汉字。

讨论汉字在跨域通讯时理論值的混乱现象，绝对不局限在简化字范围内，还包括许多异体字的取舍，即一个正字所包容的异体字范围（并／併／並；布／佈；炮／砲；匹／疋，等等）。用“繁简”来概括这一问题是不够的。从整个汉字系统上来说，理論值的非同步问题当然涵盖了繁简汉字的一对多问题，但不仅仅限于汉字的繁简问题。

三. 设计原则

在了解了讨论汉字时必备的一些概念之后，我们回到本文所要议论的主要话题。先于此，我们必须理解统一码的设计原则是如何言及文字的抽象概念的。

统一码的十条编码原则之一，就是要求字符的编码不针对具体的字形。即所谓的“字符，而非字形”原则（见表 1）。因为电子计算机所要传递或接受的“文字”必须是互相之间不可能存在误解的电子信号，是无形的概念性统一文字，而绝不是因人而异的图形性文字。对于信息通讯的核心诉求来说，文字的图形性因素甚至是干扰因素。字符（Character）指的是相当于信号的文字概念，而字形（glyph）则是图形性的、具体的文字表达。

笔者已经表明过，汉字符号系统的最小单位就是汉字单位，理想的每一个汉字单位在字义功能上必定有别于任何其他汉字单位，也就是说，它的存在和文字功能必须能区别于其他汉字单位。笔

原则	描述
通用性	Unicode标准提供了一个单一的、通用的指令系统
有效性	Unicode文本易于解析和处理
字符，非字形	**Unicode标准对字符编码，而不是对字形**
语义	字符有定义的语义
纯文本	Unicode字符为纯文本
逻辑顺序	Unicode文本以逻辑顺序保存
一致性	Unicode标准在文字中将不同语言统一为相同的字符
动态合成	重音格式能够动态合成
稳定性	字符一旦被分配，就不能再次分配，而且是固定不变的
可变换性	精确的可转换性确保Unicode标准与其他被广泛接受的标准可以相互转换

表1：10项 Unicode 设计原则（引自孙伟峰, 李德龙译《Unicode 5.0 标准》），粗体为笔者所为。

者还表明，汉字单位是个抽象的概念。不同的使用者按照自己的用字标准、用字习惯和语境对其赋予“正确的”形值和音值。这些论述表明，汉字作为符号系统的最小单位，它需要区别的不是发音上的相同或不同，也不是字形字体上是一致还是有差异。而是纯粹地以其文字功能（主要体现在文字功能上的差别）这一符号基本单位恰恰就是数码化符号的最小区别单位。也即是说做为电子文件需要表达的基本符号差异，是和汉字的基本符号单位一致的。统一码原则中所说的**字符**（Character）就相当汉字单位，这是一个不受图形约束的电磁信号。而**字形**（glyph）才是涉及汉字笔画、风格、繁简等因素的、和汉字的字形规范有关的、印刷成书面形式的具体汉字。

笔者能够理解这一认识会引起争议，汉字使用国家和地区在形成统一码的努力过程也绝不是一帆风顺，中间经过复杂艰苦的磨合，才形成今天的结果。所以对于汉字编码如何形成新的共识，毫无疑问或许也同样需要一个复杂艰苦的磨合过程。

为了对汉字编码是如何形成的过程增加感性认识，本文愿意先从早年的一些编码思路去寻找发现问题的突破口。

日本在世界上最先推出计算机汉字编码后，在中文领域的第一个汉字编码是由台湾的民间机构国字整理小组研发的**中文资讯交换码**（下称CCCII）。该编码的设计意图之一是为了向北美图书馆东亚图书目录的联机上网提供一份实用可靠的汉字目录，所以不仅是台湾汉字，还包括了大陆的简化字，以及日本使用的汉字。虽然简化字和相应的繁体字按照编码原理来说本来应该是同一个“字符”，但由于当时的编码技术所限，或更主要的是因为用户对于电脑表现各种汉字字体的要求，所以兼收并蓄各种汉字字体，包括繁简汉字和常见的异体字是不足为奇的。

即使如此，我们必须注意到中文资讯交换码对于台湾正字和大陆简化字（大陆正字）采用了在码位安排上对应的方式。简单地表达就是：台湾汉字在第一层面，大陆简化字安排在在第二层面的相应位置，与繁体字上下相应对齐。其他层面，则依次是日本的字体和异体字，也同样与第一层面的台湾正字上下对齐。虽然这种安排形式会让其他汉字使用地区觉得别扭，但是它的理论框架说明设计人员强烈地意识到不同字体字形汉字在理论上是同一个字符的思想萌芽。

大陆的第一个编码字符集 GB 2312-80（GB基本集，下称GB0）主要体现了大陆内地的汉字规范。不久之后，又为了应对古籍汉字的上机或印刷，在继 GB 2312-80之后，开发了号称和简化字对应的繁体字版 GB12345-90（GB第一辅助集，下称GB1）。GB 和台湾CCCII 的不同之处，是不仅将和简化字对应的繁体字进行编码，对非繁非简汉字也都重复编码。也就是说在数码化领域设计了两份漢字清单，一份清单包含简化字，另一份包含繁体字。和CCCII的单字切换的设计思想的第一个不同点是：GB1的设计意图是实现全汉字繁简切换。这种设计的好处是可以利用电脑的动作实现高速操作。看起来大量重复，其实是具有一定意义的合理性和便捷性的。第二个不同之处，也是尤其需要指出的是：大陆内地的繁体字（严格地应称为选用字）很多是修整、统合传统字形的新字形（广义的说亦可算作大陆规范字）[9]。换个思路，从字符和字形的立场上看，大陆内地的设计思想更偏重于字形上的规范，犹如是在规范计算机字模。事实上国家技术监督局还相继公布了一系列“信息交换用汉字字模集”（1985～）等规范文件。

笔者推测当时的编码思想只是为了体现“境内”的用字规范，所以即便GB0和GB1并非百分之百地一一对应，也似乎没有让设计者产生忧虑，GB1相对于GB0溢出103个字，被安排到汉字表的最末尾。或许出于字频考虑，溢出汉字不都是繁体字，也有简化字。现在看来这种粗糙的安排是考虑不周的。尤其是这种看重字体笔划，轻视数码文字抽象性的编码思想埋下了不小的隐患。（图1显示出GB1繁体选用字溢出的概念。

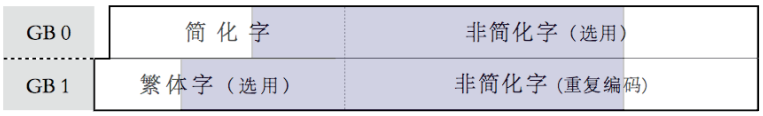

图1：从理论上说，大陆规范字形的覆盖面包括常用规范字中的简化字、非简化(选用)字、以及繁体(选用)字。 ：大陆规范字形

因为企图利用切换编码兼容繁简汉字的想法虽然在理论上是合理的，但却与大陆内地的繁体字政策相矛盾。因为大陆的繁体字原理上仅限用于古籍文献，而古籍文献对应简化字是一个单向的对应关系。多对一的繁简关系在由繁向简的转换时，是一个简单的合并动作（少数繁一对简多的情况则相反），也是电脑基本能够承担的。理论上来说，现代人写的包含简化字的文件也没有让古人阅读的可能性，换句话说，由简向繁的文件转换，理论上是不存在的。繁简汉字之间，即GB0和GB1之间虽然不能完整兼容，但因为现实当中不存在必须解决往返转换的课题，其缺陷可以忽略不计。

但是这个缺陷其实也是一个致命的漏洞，那就是没有考虑境外的中文使用环境。编码设计人员本身也许都没有意识到GB1的使用环境和台湾正字根本不同，使用者很可能单纯地将包含繁体字的GB1视为台湾汉字的可代替物。这就实际上产生了GB0和GB1必须完整兼容的要求。同时，往返双向转换就变成不可避免的课题了。笔者认为，在大陆内地汉字和台湾汉字不能顺利转换之前，其实已经存在GB0和GB1的对应障碍。这样，图2的“X”就可以说是数码汉字非完整兼容的第一原因。

笔者记得当年国内外流行语言处理机，中文的语言处理机器右上角有一个“繁简切换键”，按它一下，文件的字体就由简化字版变成繁体字版，再按一下就又魔术般地变回来。笔者也曾经抱着一线希望试着玩过（结果是可想而知的）。九十年代，笔者还从朋友那儿得到过GB版的繁体字电脑

字模（Computer font），在电脑上只需用切换字模的方法，就能让原本简化字版的文章瞬息间变成繁体字版。当然，和中文处理机编码繁简转换键同样，其结果是无法用准确一词来形容的。

繁简非一一对应的问题并不只是GB0和GB1，或者GB0和台湾汉字之间的问题。还有一部分是除了部分简化字包容了其他汉字以外，部分未简化的汉字也吸收了一些异体字[10]。这些情况往往在历史上呈现半包孕异体字，在使用习惯上常会存在歧见，故总字数就会不同，这也应视为因汉字单位的不同步造成的兼容问题。图2的“Y”大致表示数码汉字非完整兼容的第二原因。

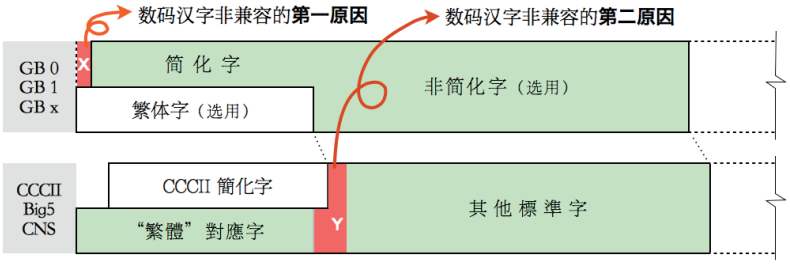

图2：“繁简”汉字的非对应，可能只包括了两岸规范汉字“非对应”的一部分。

四. 将错就错

到了统一码的时代，CCCII 的繁简对应或 GB0 对 GB1 的（即便是有问题的）对应框架已经不复存在。它的国内相应编码字符集就是 GB 13000.1-93。，由于当年人们尚未找到如何在同一个码位上使不同地区的不同字形汉字同步编码，信息业界热衷于采纳广大用户对字体的具体要求并将其以外字形式随意嵌入编码。这样，这一原本应该由电脑字模领域吸收的用户要求，不断地挤压编码。这种缺乏长远考虑和理论根据的编码框架逐渐占了上风。

号称“统一”的国际编码，尽管初衷是要统一各国的汉字表，但是仍然由于理论准备的不足，实际实现“统一”（事实上是某种意义的同步）的只是有限的一部分，这是由于统一码从起步时起就以汉字的字形相似程度的大小，严格划分了能不能统一（同步）的分界线。可是我们知道做为字符的汉字单位，理论上并不是以其字体或字形决定汉字的同一性的[11]。

如此这般，按照统一码编码框架，在第一时间就首先将简化字分离出来另辟一片码位。也就是所有的简化字都被宣布和繁体字为“不同的”汉字，只需要切换编码表就能获得繁简汉

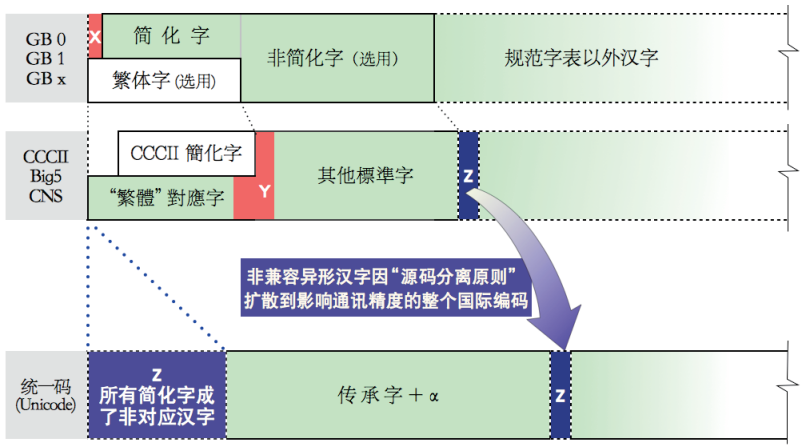

图3：非兼容的**第三个原因**在于对汉字本身抽象性，以及对统一码“字符”定义抽象性的认识不足；对汉字同一性理解的深刻分歧；及其他失误。皆造成统一码汉字在跨域通讯时发生兼容异常。

字的GB0和GB1，在统一码里成了必需转换的两套字表。而进一步加剧问题的复杂性的，是实施了源码分离原则。这一原则主张若在某个地区已经分别编码的汉字，即使十分相似，也必须分别编码。这听起来“技术性”颇强的“原则”其实直接和统一码的“字符，而非字形”编码原则相冲突的。故按笔者的看法不如说是个“错上加错”原则。（图3的“Z”大略表示统一码造成的区域间汉字兼容障碍的第三原因）。

下面列出一些字形或笔形有差异的汉字，看看在统一码中分别受到怎样的对待。

例 ①

a. 靑（U+9751）　　b. 青（U+9752）

两字在源码中分别编码，属于不认同的源码分离汉字，一般互相无法检索。

例 ②（包含例 ① 的字体）

a. 倩（U+5029）　　b. 倩（U+5029）

两字虽有笔形差异，但按照统一码认同规则属于认同汉字，能够互相检索。此类情况还有如：

骨／骨，差／差，鬼／鬼 ･･･

例 ③

a. 直（U+76F4）　　b. 直（U+76F4）

两字虽有笔形差异，但按照统一码认同规则属于认同汉字，能够互相检索。

例 ④（包含例 ③ 的字体）

a. 値（U+5024）　　b. 值（U+503C）

两字在源码中分别编码，虽然按照统一码认同规则属于认同汉字，但在源码分离原则下，成为非认同的汉字，一般互不能检索。

例 ⑤

a. 每（U+6BCE）　　b. 每（U+6BCF）

两字的笔划不同，但按照统一码认同规则属于认同汉字，但又因为在源码中分别编了码，受源码分离原则影响，成为非认同的汉字，互相无法检索。

例 ⑥（包含例 ⑤ 的字体）

a. 海 （ U+6D77）　　b. 海 （ U+6D77 ）

a、b 两字笔划数不同，但按照统一码认同规则属于认同汉字。能互相检索。

例 ⑦

a. 戶（U+6236）　b. 户（U+6237）　c. 戶（U+6238）

a和b、c三者在源码中分别编码，属于不得统合的源码分离汉字，互相之间无法检索。相同情况还有如比如：

淸（U+6DF8）　　清（U+6E05）

將（U+5C07）　　将（U+5C06）

类似上面所举出的，某些汉字认同而某些汉字不认同，某些不认同汉字作为部件的汉字却属于认

同汉字，相反一些以认同汉字作部件的汉字，虽然毫无其他不同之处却被处理成不认同汉字。如此这般的矛盾与混乱，在统一码的汉字表中俯拾皆是，不一而足。皆可说明“源码分离原则”其实是如何地“无原则”。

五. 汉字同步

本文粗略的描述了统一码中的汉字编码问题。从汉字符号理论出发来验证，比较容易看到：依据现行汉字认同规则为背景建立的编码框架，是造成各区域之间数码汉字信息交换不畅通的根本原因。

那么，目前是否存在实现汉字的真正合理统一编码的可能性呢?

从结论上说，对统一码加工改造在现行体制中是行不通的。因为统一码另外有一个重要的“稳定性”原则：“字符一旦被分配，就不能再次分配，而且是固定不变的”。尽管现在统一码的汉字编码框架思路有偏误，且经过多次增补，在汉字的取舍、编排上破绽百出。但是这毕竟是经由各国编码专家们讨论妥协的结果，不可能在一夜之间说改就改。

解决上述问题，从根源上说，还是一个对汉字本身的认识问题和现代汉字符号学研究领域相对滞后所造成的。或者说，围绕汉字规范的诸多问题，既是一个汉字系统自身的本体规划问题，也是一个汉字数码化理论的问题，同时又是一个数码汉字编码原则以及编码框架上的技术性问题。这些问题长期以来互相缠绕，纠结不清，使得人们看不清解决问题从何处着手。

笔者认为，汉字同步（汉字单位的统一规划）绝不可能是某个国家或地区单独能解决的，这是相关国家和地区共同参与，多边协作才能走得通的道路。在这场优化汉字体系的变革中，围绕中文汉字，海峡两岸文字规范部门的文字观念和政策措施将是首要关键。中文数码汉字领域的整体规划建设不跨出第一步，东亚乃至全球就难以实现稳定、精确、可靠的汉字信息传递。

注

1）参见《新英和大辞典》第6版（［日］竹林滋著, 研究社（2002））。

2）该词典将GRAPHEME解释成“字位，字素”，没有言及和汉字的关系。

3）参见佐藤喜代治等人编撰《漢字百科大事典》（明治书院, 1996, 东京）。

4）见菱沼透（1984）

5）日本《表外汉字字体表》（2000年12月8日日本文化厅国语审议会报批）的“1.前文”当中解释称“字种”为调查统计报纸杂志用字，计算使用频率时的“不同的字”。由此，我们尚无法断定“字种”一词是否具有经得起严密表达的术语性。

6）见陈明然（2014）。

7）见许长安（1992）。

8）CJK统一汉字中台湾的源码规范是 CNS 11643-200 ，本文取习惯称呼。

9）详见《第一批异体字整理表》（1955年12月）《印刷通用汉字字形表》（1965年1月）。

10）详见《第一批异体字整理表》（1955年12月）。

11）见朱一星(2013)。

参考文献

R.R.K.哈特曼, F.C.斯托克 1981 《语言与语言学词典》, 上海辞书出版社

菱沼透 1984 中国の標準字体と日本の通用字体, 《日本語学》3卷3号 , 明治书院, pp.32-40

柯少斋译 1987 美国国会图书馆电脑编目, 《图书馆学与资讯科学》13期, pp.210-223

许长安 1992 海峡两岸用字比较《语文建设》1992年1期, pp.13-18

张鼎锺 1993 中文资讯交换码与中文图书资料自动化之回顾 , 《鼎鐘文集》, pp.43-51

黄克东 1994 “书同文” 应向前迈进一步 , 《中文信息》1994年第5期, pp.3-10

冯志伟 2000 论语言文字的地位规划和本体规划, 《中国语文》2000年04期

苏培成 2004 重新审视简化字 , 史定国主编《简化字研究》, 商务印书馆, pp.63-81

李 鎏 2006 从学术观点看「正体字」与「简化字」, 《语文建设通讯》99期

杨宝忠 2008 ISO-10646 国际编码字符集存在的问题, 第五届两岸四地中文数字化论坛, 合肥

松冈荣志 2010 《漢字・七つの物語 中国の文字改革一〇〇年》, 三省堂

王 宁, 王立军 2011 当前汉字规范的一个重要问题——兼谈《通用规范汉字表》对这些问题的处理, 《汉语教学与研究》第一辑, pp.33-42

苏培成 2011 论《通用规范汉字表》的修订和完善, 《汉语教学与研究》第一辑, pp.33-42

小林龙生 2011 《ユニコード戦記 —文字符号の国際標準化バトル》, 东京电机大学出版局

陈明然 2014 汉字规範和汉字信息处理技術, 《语文建设通讯》第105期, pp.29-31

朱一星 2013 如何界定汉字的理论单位, 京都外国语大学《研究论丛》No.81, pp.129-139

相关資料

孙伟峰, 李德龙译 2009 Unicode 协会著 《Unicode 5.0 标准》, 清华大学出版社

Unicode 8.0.0 2015 , 网页: http://unicode.org/versions/Unicode8.0.0/

Unicode ideographic variation database（UTS #37）, 网页: http://unicode.org/reports/tr37/